Detection and Estimation with Compressive Measurements

Mark A. Davenport, Michael B. Wakin, and Richard G. Baraniuk[*]

Rice University
Department of Electrical and Computer Engineering

Technical Report TREE 0610

November 1, 2006 (Updated January 24, 2007)

Abstract

The recently introduced theory of compressed sensing enables the reconstruction of sparse or compressible signals from a small set of nonadaptive, linear measurements. If properly chosen, the number of measurements can be much smaller than the number of Nyquist rate samples. Interestingly, it has been shown that *random projections* are a satisfactory measurement scheme. This has inspired the design of physical systems that directly implement similar measurement schemes. However, despite the intense focus on the reconstruction of signals, many (if not most) signal processing problems do not require a full reconstruction of the signal – we are often interested only in solving some sort of *detection* problem or in the *estimation* of some function of the data. In this report, we show that the compressed sensing framework is useful for a wide range of statistical inference tasks. In particular, we demonstrate how to solve a variety of signal detection and estimation problems given the measurements without ever reconstructing the signals themselves. We provide theoretical bounds along with experimental results.

1 Introduction

Over the past decades the amount of data generated by sensing systems has grown from a trickle to a torrent. This has stimulated a great deal of research in the fields of compression and coding, which enable compact storage and rapid transmission of large amounts of information. Compression is possible because we often have considerable a priori information about the signals of interest. For example, many signals are known to have a *sparse* representation in some transform basis (Fourier, DCT, wavelets, etc.) and can be expressed or approximated using a linear combination of only a small set of basis vectors.

[*]MAD and RGB are with the Department of Electrical and Computer Engineering, Rice University, Houston, TX. Email: {md, richb}@rice.edu. MBW is with the Department of Applied & Computational Mathematics, California Institute of Technology, Pasadena, CA. Email: wakin@acm.caltech.edu.

This research was supported by ONR grants N00014-06-1-0769 and N00014-06-1-0829; AFOSR grant FA9550-04-0148; DARPA grants N66001-06-1-2011 and N00014-06-1-0610; NSF grants CCF-0431150, CNS-0435425, and CNS-0520280; and the Texas Instruments Leadership University Program.

Web: dsp.rice.edu/cs

The traditional approach to compressing a sparse signal is transform coding – we compute its transform coefficients and then store or transmit the few large coefficients along with their locations. This is an inherently wasteful process (in terms of both sampling rate and computational complexity), since it forces the sensor to acquire and process the entire signal even though an exact representation is not ultimately required. In response, a new framework for simultaneous sensing and compression has developed recently under the rubric of *Compressed Sensing* (CS). CS enables a potentially large reduction in the sampling and computation costs at a sensor. CS builds on the work of Candès, Romberg, and Tao [1] and Donoho [2], who showed that a signal having a sparse representation in some basis can be reconstructed from a small set of nonadaptive, linear measurements. Briefly, this is accomplished by generalizing the notion of a measurement or sample to mean the application of a linear functional to the data. We can represent this measurement process in terms of matrix multiplication. In [1, 2] conditions upon this matrix are given that are sufficient to ensure that we can stably recover the original signal using a tractable algorithm. Interestingly, it can be shown that with high probability, a matrix drawn at random will meet these conditions.

CS has many promising applications in signal acquisition, compression, medical imaging, and sensor networks [1–9]; the random nature of the measurement matrices makes it a particularly intriguing *universal* measurement scheme for settings in which the basis in which the signal is sparse is unknown by the encoder or multi-signal settings in which distributed, collaborative compression can be difficult to coordinate across multiple sensors. This has inspired much interest in developing real-time systems that implement the kind of random measurement techniques prescribed by the CS theory [10–12]. Along with research into new measurement systems, a variety of reconstruction algorithms have been proposed [1, 2, 13], all of which involve some kind of iterative optimization procedure, and thus are computationally expensive for long signals with complexity typically polynomial in the signal length.

While the CS literature has focused almost exclusively on problems in signal reconstruction or approximation, this is frequently not necessary. For instance, in many signal processing applications (including most communications and many radar systems), signals are acquired only for the purpose of making a detection or classification decision. Our aim in this paper is to show that the CS framework is useful for a much wider range of statistical inference tasks. Tasks such as detection do not require a reconstruction of the signal, but only require estimates of the relevant *sufficient statistics* for the problem at hand. Our key finding is that in many cases it is possible to directly extract these statistics from a small number of random projections without ever reconstructing the signal. This work expands on the previous work on detection using compressive measurements in [14, 15]. We also build upon and draw parallels between the large body of *sketching* literature where random or pseudo-random measurement techniques have long been used to estimate various quantities of large data streams. In particular, our results on the estimation of linear functions of the data parallel those of [16].

This report is organized as follows. Section 2 provides background on CS and dimensionality reduction. In Section 3 we analyze some simple signal detection and classification problems, and demonstrate the effectiveness of the compressive detector on some realistic problems. Section 4 considers the problem of estimating linear functions of the original signal, and studies the performance of a simple compressive estimator in this setting. Finally, Section 5 describes the relationship between our work and previous work on sketching algorithms for large data streams, and Section 6 concludes with directions for future work.

2 Concentration of Measure and Compressed Sensing

Let $x \in \mathbb{R}^N$ be a signal and let the matrix $\Psi := [\psi_1, \psi_2, \ldots, \psi_Z]$ be a basis for $\mathbb{R}^N$. Let Σ_K denote the set of all K-*sparse* signals, by which we mean that any $x \in \Sigma_K$ can be represented as a linear combination of K vectors from Ψ. We then consider an $M \times N$ measurement matrix Φ. A somewhat surprising result of [2, 17] is that there exist matrices Φ such that every $x \in \Sigma_K$ can be recovered *exactly* from $y = \Phi x$ using a simple and computationally tractable algorithm (ℓ_1 minimization), provided that $M = O(K \log(N/K)) \ll N$. In fact, one can show that we can obtain such a Φ by simply picking one at random, or equivalently, by selecting a Φ that projects our data onto a random M-dimensional subspace.

It may seem improbable that we can recover the original signal x *exactly* from the measurements y, after all, such inverse problems are generally ill-posed whenever $M < N$. However, CS recovery algorithms exploit the additional assumption of *sparsity* in Ψ to identify the correct signal x from an uncountable number of possibilities. Specifically, CS recovery is possible because with high probability, a randomly chosen Φ results in a "stable" embedding of the sparse signal set Σ_K in $\mathbb{R}^M$ (any two well-separated sparse signals in $\mathbb{R}^N$ remain well-separated upon projection to $\mathbb{R}^M$). One way of characterizing this stability is known as the *Restricted Isometry Property* (RIP); we say Φ has RIP of order K if for every $x \in \Sigma_K$,

$$(1-\epsilon)\sqrt{\frac{M}{N}} \leq \frac{\|\Phi x\|_2}{\|x\|_2} \leq (1+\epsilon)\sqrt{\frac{M}{N}}. \tag{1}$$

For a thorough review of the various results illustrating the role of the RIP in producing a stable embedding of Σ_K, see [18]. One might initially question the existence of matrices satisfying (1), but in fact a random orthoprojector[1] from $\mathbb{R}^N$ to $\mathbb{R}^M$ can be shown to meet the RIP of order K (with respect to any fixed sparsity basis Ψ) with high probability if $M = O(K \log(N/K))$.

In a recent paper [19], we have identified a fundamental connection between the random matrix constructions of CS and the Johnson-Lindenstrauss (JL) lemma [20–22], which concerns the stable embedding of a finite set of points under a random dimensionality-reducing projection.

Lemma 2.1 [Johnson-Lindenstrauss] *Let S be a finite collection of points in $\mathbb{R}^N$. Fix $0 < \epsilon < 1$ and $\beta > 0$. Let Φ be a random orthoprojector from $\mathbb{R}^N$ to $\mathbb{R}^M$ with $M < N$ and*

$$M \geq \left(\frac{4+2\beta}{\epsilon^2/2 - \epsilon^3/3}\right) \ln(|S|), \tag{2}$$

where $|S|$ denotes the number of elements of S. With probability exceeding $1-(|S|)^{-\beta}$, the following statement holds: For every $s_i, s_j \in S$, $i \neq j$,

$$(1-\epsilon)\sqrt{\frac{M}{N}} \leq \frac{\|\Phi s_i - \Phi s_j\|_2}{\|s_i - s_j\|_2} \leq (1+\epsilon)\sqrt{\frac{M}{N}}. \tag{3}$$

[1] By an orthoprojector, we mean an orthogonal projection from $\mathbb{R}^N$ to $\mathbb{R}^M$ that can be expressed as an $M \times N$ matrix Φ with orthonormal rows. A random orthoprojector may be constructed, for example, by running the Gram-Schmidt process on M random length-N vectors having i.i.d. Gaussian entries (assuming the vectors are linearly independent). We note also that other formulations of the RIP and JL lemma pertain to matrices renormalized by $\sqrt{N/M}$ to ensure that $\|\Phi x\|_2 \approx (1 \pm \epsilon)\|x\|_2$. However, we find it useful in this paper to work with random orthoprojectors and the resulting "compaction" by $\sqrt{M/N}$.

Without going into complete detail, let us briefly describe how one proves the JL lemma using such random matrices. First, show that for any $x \in \mathbb{R}^N$, the random variable $\|\Phi x\|_2^2$ has expected value $(M/N)\|x\|_2^2$; that is,

$$\mathbb{E}(\|\Phi x\|_2^2) = (M/N)\,\|x\|_2^2\,. \tag{4}$$

Second, show that for any $x \in \mathbb{R}^N$, the random variable $\|\Phi x\|_2^2$ is strongly concentrated about its expected value; that is,

$$\Pr\left(\left|\|\Phi x\|_2^2 - (M/N)\,\|x\|_2^2\right| \geq \epsilon(M/N)\,\|x\|_2^2\right) \leq 2e^{-\frac{M}{2}(\epsilon^2/2-\epsilon^3/3)}, \quad 0 < \epsilon < 1, \tag{5}$$

where the probability is taken over all $M \times N$ matrices Φ [22]. To prove the JL lemma one applies (5) to the $\binom{|S|}{2}$ vectors corresponding to all possible interpoint distances. (In the following, it will be useful to note that if $\delta \in (0,1)$ is a fixed value and $\epsilon < C_\delta/\sqrt{M}$, where $C_\delta = \sqrt{12\log(2/\delta)}$, then the right-hand side of (5) is less than δ.)

Let us now mention some examples of distributions that satisfy our concentration inequality (5). As mentioned above, random orthoprojectors will, but several other types of random matrices also satisfy (5). The simplest examples are the $M \times N$ random matrices Φ whose entries $\phi_{i,j}$ are independent realizations of Gaussian random variables

$$\phi_{i,j} \sim \mathcal{N}\left(0, \frac{1}{N}\right). \tag{6}$$

One can also use matrices where the entries are independent realizations of ± Bernoulli random variables

$$\phi_{i,j} := \begin{cases} +\frac{1}{\sqrt{N}} & \text{with probability } \frac{1}{2}, \\ -\frac{1}{\sqrt{N}} & \text{with probability } \frac{1}{2}, \end{cases} \tag{7}$$

and a related distribution yields the matrices

$$\phi_{i,j} := \begin{cases} +\sqrt{\frac{3}{N}} & \text{with probability } \frac{1}{6}, \\ 0 & \text{with probability } \frac{2}{3}, \\ -\sqrt{\frac{3}{N}} & \text{with probability } \frac{1}{6}. \end{cases} \tag{8}$$

As discussed in [22], the verification of (5) in the Gaussian case is elementary (using tail bounds for Gamma random variables), and the other two cases are verified in [22] using a straightforward relation to the moments of Gaussian random variables.

At first glance there are several apparent differences between CS (which deals with embedding an *uncountable* point set and correctly identifying a signal from its projections) and the JL lemma (which deals only with embedding a *finite* number of points and makes no mention of signal recovery). However, for the purpose of ensuring a stable CS embedding, Σ_K (because of its limited complexity) can be characterized in terms of a finite number of point samples. By applying the JL lemma only to these points we can deduce the RIP for all of the remaining points on Σ_K, which in turn permits stable CS signal recovery [19]. A similar technique has recently been used to demonstrate that random projections also provide a stable embedding of smooth manifolds [23]. We will now make further use of this connection in the following sections to aid in our analysis of the performance of the use of compressive measurements in a number of detection and estimation problems.

3 Compressed Detection

3.1 The Traditional Detector

We begin by examining the simplest of detection problems. We would like to distinguish between $\mathcal{H}_0$ and $\mathcal{H}_1$:

$$\begin{aligned}\mathcal{H}_0 &: x = n\\ \mathcal{H}_1 &: x = s + n\end{aligned}$$

where $s \in \mathbb{R}^N$ is a known signal, and $n \sim \mathcal{N}(0, \sigma^2 I_N)$ is i.i.d. Gaussian noise. Next, let

$$\begin{aligned}P_F =& \Pr(\mathcal{H}_1 \text{ chosen when } \mathcal{H}_0 \text{ true) and}\\ P_D =& \Pr(\mathcal{H}_1 \text{ chosen when } \mathcal{H}_1 \text{ true)}\end{aligned}$$

denote the *false alarm rate* and the *detection rate* respectively. The *Neyman-Pearson* (NP) detector is the decision rule that maximizes P_D subject to the constraint that $P_F \leq \alpha$. It is not difficult to show (see [24, 25], for example) that the NP-optimal decision rule is the *likelihood ratio test*:

$$\Lambda(x) = \frac{f_1(x)}{f_0(x)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta$$

where η is chosen such that

$$P_F = \int\limits_{\Lambda(x)>\eta} f_0(x)dx = \alpha.$$

For our hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, we have

$$f_0(x) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{\|x\|_2^2}{2\sigma^2}\right) \qquad \text{and} \qquad f_1(x) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{\|x-s\|_2^2}{2\sigma^2}\right).$$

By taking a logarithm we obtain an equivalent test that simplifies to

$$t := \langle x, s\rangle \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \sigma^2 \log(\eta) + \frac{1}{2}\|s\|_2^2 := \gamma.$$

It can be shown that t is a *sufficient statistic* for our detection problem, and thus t contains all the information required to determine between $\mathcal{H}_0$ and $\mathcal{H}_1$.

Returning to our detection problem, we must now set γ to achieve the desired performance. It is easy to show that

$$\begin{aligned}t \sim& \mathcal{N}(0, \sigma^2\|s\|_2^2) \quad \text{under } \mathcal{H}_0\\ t \sim& \mathcal{N}(\|s\|_2^2, \sigma^2\|s\|_2^2) \quad \text{under } \mathcal{H}_1.\end{aligned}$$

Thus we have

$$\begin{aligned}P_F =& P(t > \gamma|\mathcal{H}_0) = Q\left(\frac{\gamma}{\sigma\|s\|_2}\right),\\ P_D =& P(t < \gamma|\mathcal{H}_1) = Q\left(\frac{\gamma - \|s\|_2^2}{\sigma\|s\|_2}\right),\end{aligned}$$

where

$$Q(z) = \int_z^\infty \exp\left(-\frac{u^2}{2}\right) du.$$

To determine the appropriate value for γ, we set $P_F = \alpha$, which yields

$$\gamma = \sigma \left\|s\right\|_2 Q^{-1}(\alpha)$$

resulting in

$$P_D(\alpha) = Q\left(Q^{-1}(\alpha) - \frac{\|s\|_2}{\sigma}\right). \tag{9}$$

If we define

$$\text{SNR} := \frac{\|s\|_2^2}{\sigma^2} \tag{10}$$

then we can rewrite (9) as

$$P_D(\alpha) = Q\left(Q^{-1}(\alpha) - \sqrt{\text{SNR}}\right). \tag{11}$$

3.2 The Compressive Detector

We now consider the same detection problem as in Section 3.1, but instead of observing x we observe $y = \Phi x$ where $\Phi \in \mathbb{R}^{M \times N}$, $M \leq N$. Our problem now is to distinguish between $\widetilde{\mathcal{H}_0}$ and $\widetilde{\mathcal{H}_1}$:

$$\begin{aligned} \widetilde{\mathcal{H}_0} &: y = \Phi n \\ \widetilde{\mathcal{H}_1} &: y = \Phi(s + n) \end{aligned}$$

where as before, $s \in \mathbb{R}^N$ is a known signal and $n \sim \mathcal{N}(0, \sigma^2 I_N)$ is i.i.d. Gaussian noise, and Φ is a fixed and known measurement matrix. In this case we have

$$f_0(y) = \frac{\exp\left(-\frac{1}{2}y^T(\sigma^2\Phi\Phi^T)^{-1}y\right)}{|\sigma^2\Phi\Phi^T|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}} \qquad \text{and} \qquad f_1(y) = \frac{\exp\left(-\frac{1}{2}(y-\Phi s)^T(\sigma^2\Phi\Phi^T)^{-1}(y-\Phi s)\right)}{|\sigma^2\Phi\Phi^T|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}}$$

Again, by taking a logarithm we obtain an equivalent test that simplifies to

$$y^T(\Phi\Phi^T)^{-1}\Phi s \underset{\widetilde{\mathcal{H}_0}}{\overset{\widetilde{\mathcal{H}_1}}{\gtrless}} \sigma^2 \log(\eta) + \frac{1}{2}s^T\Phi^T(\Phi\Phi^T)^{-1}\Phi s := \gamma.$$

We now define the compressive detector:

$$\widetilde{t} := y^T(\Phi\Phi^T)^{-1}\Phi s. \tag{12}$$

As before, it is easy to show that

$$\begin{aligned} \widetilde{t} &\sim \mathcal{N}(0, \sigma^2 s^T\Phi^T(\Phi\Phi^T)^{-1}\Phi s) \quad \text{under } \widetilde{\mathcal{H}_0} \\ \widetilde{t} &\sim \mathcal{N}(s^T\Phi^T(\Phi\Phi^T)^{-1}\Phi s, \sigma^2 s^T\Phi^T(\Phi\Phi^T)^{-1}\Phi s) \quad \text{under } \widetilde{\mathcal{H}_1}. \end{aligned}$$

Thus we have

$$\begin{aligned} P_F =&P(t > \gamma|\mathcal{H}_0) = Q\left(\frac{\gamma}{\sigma\sqrt{s^T\Phi^T(\Phi\Phi^T)^{-1}\Phi s}}\right) \\ P_D =&P(t < \gamma|\mathcal{H}_1) = Q\left(\frac{\gamma - s^T\Phi^T(\Phi\Phi^T)^{-1}\Phi s}{\sigma\sqrt{s^T\Phi^T(\Phi\Phi^T)^{-1}\Phi s}}\right). \end{aligned}$$

To set the threshold, we set $P_F = \alpha$, and thus

$$\gamma = \sigma\sqrt{s^T\Phi^T(\Phi\Phi^T)^{-1}\Phi s}Q^{-1}(\alpha)$$

resulting in

$$P_D(\alpha) = Q\left(Q^{-1}(\alpha) - \frac{\sqrt{s^T\Phi^T(\Phi\Phi^T)^{-1}\Phi s}}{\sigma}\right). \tag{13}$$

At this point it is worth considering the special case where Φ is an orthoprojector, in which case $\Phi\Phi^T = I_M$, (12) reduces to

$$\widetilde{t} = \langle y, \Phi s\rangle, \tag{14}$$

and we can bound the performance of the compressive detector as follows.

Theorem 3.1 *Let Φ be an $M \times N$ random orthoprojector. Then with probability at least $1-\delta$:*

$$Q\left(Q^{-1}(\alpha) - (1-\epsilon)\sqrt{M/N}\sqrt{SNR}\right) \le P_D(\alpha) \le Q\left(Q^{-1}(\alpha) - (1+\epsilon)\sqrt{M/N}\sqrt{SNR}\right), \tag{15}$$

where $\epsilon < C_\delta/\sqrt{M}$.

Proof: Since Φ is an orthoprojector, we can rewrite (13) as

$$P_D(\alpha) = Q\left(Q^{-1}(\alpha) - \frac{\|\Phi s\|_2}{\sigma}\right). \tag{16}$$

From (5), if Φ is an orthoprojector selected at random, then with probability at least $1-\delta$

$$(1-\epsilon)\sqrt{M/N}\,\|s\|_2 \le \|\Phi s\|_2 \le (1+\epsilon)\sqrt{M/N}\,\|s\|_2\,, \tag{17}$$

where $\epsilon < C_\delta/\sqrt{M}$. Combining (16) and (17), and recalling the definition of the SNR from (10) the result follows. □

This tells us in a precise way how much information we lose by using random projections rather than the signal samples themselves, not in terms of our ability to reconstruct the signal as is typically addressed in CS, but in terms of our ability to solve our detection problem. Specifically, for typical values of ϵ, with high probability

$$P_D(\alpha) \approx Q\left(Q^{-1}(\alpha) - \sqrt{M/N}\sqrt{\mathrm{SNR}}\right), \tag{18}$$

which increases the miss probability by an amount determined by the SNR and the ratio of M to N. Note that it would certainly be possible to design Φ that would not reduce our ability to detect s at all – in particular, if s lies in the row span of Φ where Φ is an orthoprojector we have that

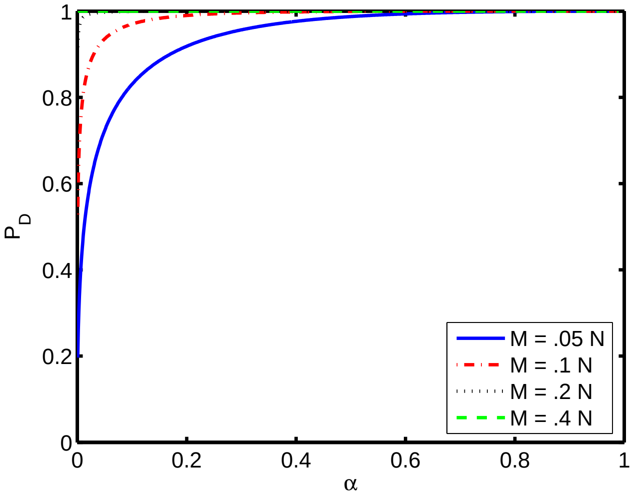

Figure 1: *Effect of M on $P_D(\alpha)$ (SNR = 20dB): For $M = .05N$ we see some degradation in performance. For $M = .1N$ we have a significant improvement in performance. By $M = .2N$ we are already very close to the same performance as that of a detector having access to the signal samples themselves.*

$\|\Phi s\|_2 = \|s\|_2$ and thus we lose nothing in terms of our ability to detect s. This makes sense since the traditional detector can be represented as a linear combination of the rows of Φ, and thus Φ really is just implementing the traditional detector for s.

In many settings, however, we might be trying to build systems that are useful in the CS setting, and thus are not able to tailor Φ to the specific s we wish to detect. However, what we lose in accuracy we gain in universality: we can provide probabilistic performance guarantees *regardless* of what s turns out to be. Thus, we can still build systems capable of detecting a signal s that is *unknown* at the time we are building the system, without simply resorting to sampling the entire signal.

Thus, we now return to the case where Φ is a matrix drawn at random. First, we examine how M affects the performance of the compressive detector. As described above, decreasing M does cause a degradation in performance. However, as illustrated in Figure 2, in certain cases (relatively high SNR; 20 dB in this example) the compressive detector can perform almost as well as the traditional detector with a small fraction of the number of measurements required by traditional detection. Furthermore, we can characterize the rate at which P_D approaches 1 as we increase M by finding a lower bound on P_D. Towards this end, in [26] we find the bound

$$Q(z) \leq \frac{1}{2}\left(1 - \sqrt{1 - e^{-z^2/2}}\right). \tag{19}$$

From this we can obtain the (very loose) bound

$$Q(z) \leq \frac{e^{-z^2/2}}{2}, \tag{20}$$

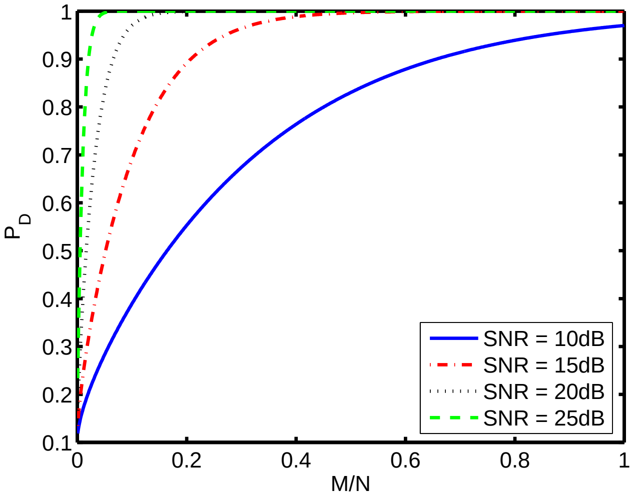

Figure 2: *Effect of M on P_D at several different SNR levels ($\alpha = .1$): Note that the rate at which P_D approaches 1 is exponentially fast, but depends on the SNR. The compressive detector can perform extremely well on just a few measurements when the SNR is high.*

which allows us to find a simple lower bound on P_D as follows: let $C_1 = (1-\epsilon)\mathrm{SNR}$, then:

$$\begin{aligned} P_D(\alpha) &\geq Q\left(Q^{-1}(\alpha) - \sqrt{M/N}\sqrt{C_1}\right) \\ &= 1 - Q\left(\sqrt{M/N}\sqrt{C_1} - Q^{-1}(\alpha)\right) \\ &\geq 1 - \frac{1}{2}e^{-\frac{C_1 M/N - 2\sqrt{C_1 M/N} + (Q^{-1}(\alpha))^2}{2}}. \end{aligned}$$

Thus, if we let

$$C_2 = \frac{1}{2}e^{-Q^{-1}(\alpha)\left(\frac{Q^{-1}(\alpha)}{2} - \sqrt{C_1}\right)} \tag{21}$$

then for all $M < N$ we obtain

$$P_D(\alpha) > 1 - C_2 e^{-\frac{C_1}{2N}M}. \tag{22}$$

Thus, for a fixed SNR and signal length, the detection probability approaches 1 exponentially fast as we increase the number of measurements. This is experimentally confirmed in Figure 2, which plots the actual performance for a range of SNRs, and in which we see that the detection probability does indeed approach 1 exponentially fast. However, we again note that in practice this rate can be significantly affected by the SNR, which determines the constants in the bound of (22).

We also note that while our discussion has been limited to Φ that are orthoprojectors, in practice this is not necessary. In Figure 3 we have plotted the ROC curve for our detector as predicted by our above analysis for orthoprojectors along with the median ROC curve (over 100 trials) for three classes of random matrices: random orthoprojectors, matrices with independent Gaussian entries, and matrices with independent Bernoulli entries. We observe almost no difference among these three methods, which all perform very much as expected.

Finally, we close by noting that for any given instance of Φ, its ROC curve may be better or worse than the curve predicted by (18). However, with high probability it is tightly concentrated around

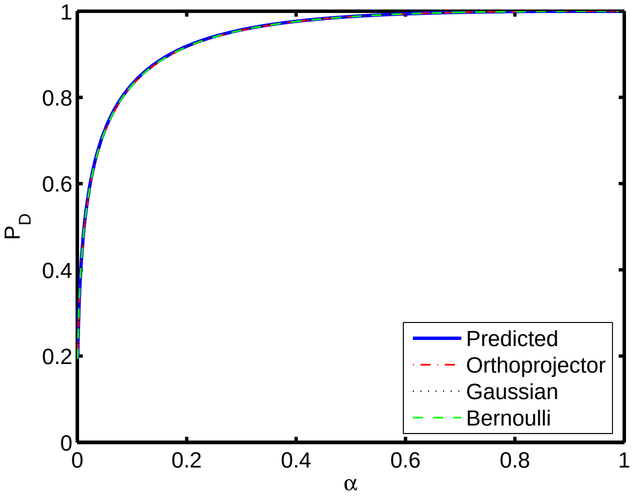

Figure 3: *Average ROC curves for different methods of generating* Φ *(SNR = 20 dB,* $M = .05N$*): The methods compared are random orthoprojectors, matrices with independent Gaussian entries, and matrices with independent Bernoulli entries. The averages are obtained by taking the median ROC curve over 100 independent trials. All agree well with the predicted ROC.*

the expected performance curve. Figure 4 illustrates this for the case where Φ has independent Gaussian entries, $M = .05N$ and $N = 10^4$. The predicted ROC curve is illustrated along with curves displaying the upper and lower bounds found by drawing 100 matrices and plotting the best and worst ROCs obtained among these. We see that our performance is never significantly different from what we expect. Furthermore, we have also observed that these bounds grow significantly tighter as we increase N – so for large problems the difference between the predicted and actual curves will be insignificant.

3.3 Compressed Classification

We can easily generalize the setting of Section 3.2 to the case where we have more than two hypotheses. In particular, suppose we have a set S of $|S|$ signals that could be present, and we would like to select between the corresponding hypotheses:

$$\widetilde{\mathcal{H}_i} \ : \ y = \Phi(s_i + n),$$

for $i = 1, 2, \ldots, |S|$, where each $s_i \in S$ is one of our known signals and as before, $n \sim \mathcal{N}(0, \sigma^2 I_N)$ is i.i.d. Gaussian noise and Φ is a known $M \times N$ measurement matrix.

In the case where Φ is an orthoprojector, Φn remains uncorrelated, and thus it is straightforward to show (see [24,25], for example), in the case where each hypothesis is equally likely, the classifier with minimum probability of error is to select $\widetilde{\mathcal{H}_i}$ where

$$\|y - \Phi s_i\|_2 < \|y - \Phi s_j\|_2$$

for all $j \neq i$. Thus, the optimal classifier depends only on the distances from y to the $|S|$ possible s_i.

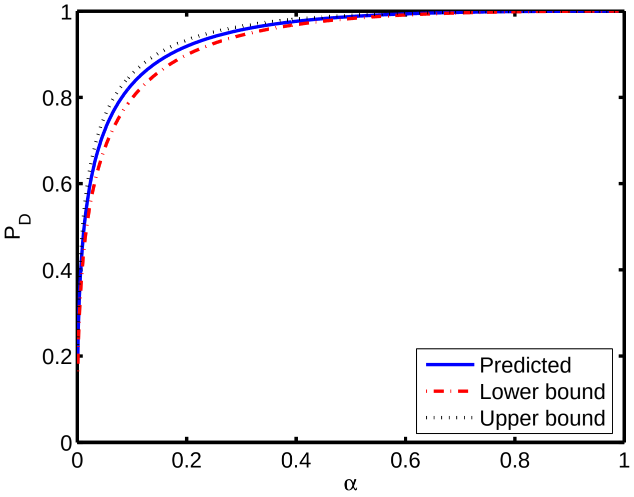

Figure 4: *Concentration of ROC curves for random Φ near the expected ROC (SNR = 20 dB, $M = .05N$, $N = 10^4$): In this example Φ has independent Gaussian entries. The predicted ROC curve is illustrated along with curves displaying the upper and lower bounds found by drawing 100 matrices and plotting the best and worst ROCs obtained among these. Our performance is never significantly different from the expected performance.*

While in general it is difficult to find analytical expressions for the probability of error in these kinds of classification settings, we can gain some intuition about the effect of the projection Φ on the performance of our classifier as follows.

Theorem 3.2 *Let Φ be an $M \times N$ random orthoprojector, and let S be a set of $|S|$ signals. Then with probability at least $1 - \delta$,*

$$(1-\epsilon)\frac{M}{N} \leq \frac{\|\Phi(x - s_i)\|_2^2}{\|x - s_i\|_2^2} \leq (1+\epsilon)\frac{M}{N}, \tag{23}$$

holds simultaneously for $i = 1, 2, \ldots, |S|$, where $\epsilon < \sqrt{(12\log(2|S|/\delta))/M}$.

Proof: By applying the union bound to (5) (in essentially the same fashion as one would to prove the JL lemma), we get that with probability at least $1 - |S|\delta'$, the following holds simultaneously for $i = 1, 2, \ldots, |S|$:

$$(1-\epsilon)\frac{M}{N} \leq \frac{\|\Phi(x - s_i)\|_2^2}{\|x - s_i\|_2^2} \leq (1+\epsilon)\frac{M}{N}, \tag{24}$$

where $\epsilon < \sqrt{(12\log(2/\delta'))/M}$. By setting $\delta = \delta'/|S|$ the result follows. □

While this may be viewed as nothing more than a simple application of the JL lemma, we emphasize in our treatment that the distances between x and the s_i shrink (approximately uniformly over all i) by a factor of $\sqrt{M/N}$, and this compaction cannot simply be ignored away by renormalizing Φ. Thus, the distances are not *preserved*, they are *uniformly shrunken*. The importance of this distinction is that if $y = \Phi(x + n)$, then because projection to a lower dimensional space shortens distances, the effect of the noise can be amplified. In particular, if the noise has covariance matrix $\sigma^2 I_N$, then if Φ is an orthoprojector, Φn has covaraince matrix $\sigma^2 \Phi\Phi^T = \sigma^2 I_M$. Thus, a small amount of noise that may not impact our classifier in the original N-dimensional space can become more significant after applying Φ. Clearly, rescaling Φ to avoid this compaction simply rescales the variance of the noise as well. Thus, just as in detection, the probability of error of our classifier will increase upon projection to a lower-dimensional space in a way that depends on the SNR.

4 Compressed Estimation

While many signal processing problems can be reduced to a detection or classification setting, in some cases we cannot reduce our task to selecting among a finite set of hypotheses. For example, we may often be interested in *estimating* some function of the data. In this section we will focus on estimating a *linear* function of the data from compressive measurements. Specifically, suppose that we observe $y = \Phi x$ and wish to estimate $\langle s, x \rangle$ from the measurements y. In the case where Φ is an orthoprojector, a natural estimator is essentially the same as the compressive deterctor (appropriately reweighted to account for the expected compaction of distances by $\sqrt{M/N}$). Specifically, suppose we have a set S of $|S|$ linear functions we would like to estimate from y, and consider the estimator:

$$q_i(y) = \frac{N}{M} \langle y, \Phi s_i \rangle , \tag{25}$$

for $i = 1, 2, \dots, |S|$.

Once again, we can again use the concentration of measure inequality of (5) to analyze the performance of this estimator as follows.

Theorem 4.1 *Let Φ be an $M \times N$ random matrix with entries drawn according to a distribution satisfying the concentration of measure inequality (5). Let S be a set of $|S|$ points in $\mathbb{R}^N$, $s_1, s_2, \dots s_{|S|}$. Then with probability at least $1 - \delta$,*

$$\left| \frac{N}{M} \langle \Phi x, \Phi s_i \rangle - \langle x, s_i \rangle \right| \leq \kappa_\delta \frac{\|x\|_2 \, \|s_i\|_2}{\sqrt{M}} \tag{26}$$

simultaneously for $i = 1, 2, \dots, |S|$, where $\kappa_\delta = 2\sqrt{12 \log(\frac{4|S|+2}{\delta})}$.

Proof: We first note that the inequality (26) holds trivially if either $\|x\|_2 = 0$ or $\|s_i\|_2 = 0$, and so we proceed under the assumption that $\|x\|_2 \neq 0$ and $\|s_i\|_2 \neq 0$. We now define

$$u = \frac{x}{\|x\|_2} \qquad \text{and} \qquad v_i = \frac{s_i}{\|s_i\|_2}. \tag{27}$$

Note that (for all i) $\|v_i\|_2 = \|u\|_2 = 1$, and thus by applying the union bound to (5) we obtain that with probability at least $1 - \delta$ the all three of the following inequalities hold for all i simultaneously:

$$(1-\epsilon)\frac{M}{N} \leq \|\Phi u\|_2^2 \leq (1+\epsilon)\frac{M}{N}, \tag{28}$$

$$(1-\epsilon)\frac{M}{N} \leq \|\Phi v_i\|_2^2 \leq (1+\epsilon)\frac{M}{N}, \tag{29}$$

$$(1-\epsilon)\frac{M}{N} \leq \frac{\|\Phi(u - v_i)\|_2^2}{\|u - v_i\|_2^2} \leq (1+\epsilon)\frac{M}{N}, \tag{30}$$

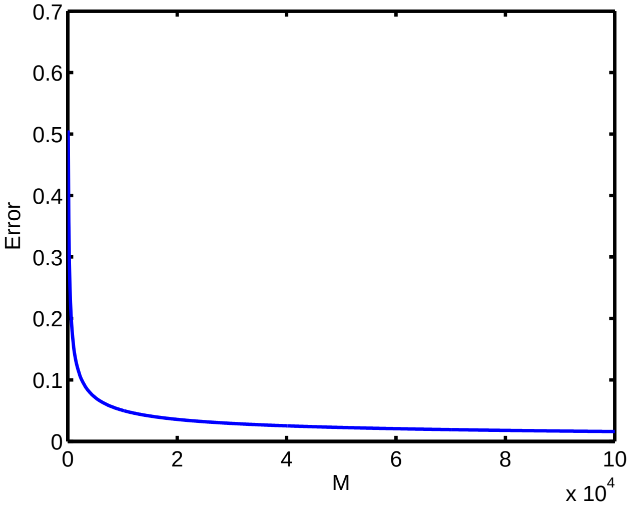

Figure 5: *Effect of M on the estimation error ($N = 10^5$): Predicted bound on error as a function of M. Note that this decays much more slowly than the probability of error for detection.*

where $\epsilon < \frac{\kappa_\delta}{2\sqrt{M}}$. Next, we observe that for all i,

$$(M/N)(1-\epsilon)\left(2 - 2\langle u, v_i\rangle\right) \leq \|\Phi u\|_2^2 - 2\langle \Phi u, \Phi v_i\rangle + \|\Phi v_i\|_2^2 \leq 2(M/N)(1+\epsilon) - 2\langle \Phi u, \Phi v_i\rangle$$

where the left-hand inequality follows from the left-hand inequality of (30) and the fact that for all i, $\|v_i\|_2 = \|u\|_2 = 1$, and the right-hand inequality follows from (28) and (29). Rearranging, we obtain:

$$\begin{aligned}(N/M)\langle \Phi u, \Phi v_i\rangle - \langle u, v_i\rangle &\leq (1+\epsilon) - (1-\epsilon) - \epsilon\langle u, v_i\rangle \\ &= (2 - \langle u, v_i\rangle)\epsilon \\ &\leq 2\epsilon,\end{aligned}$$

Proceeding similarly using the right-hand side of inequality (30) we obtain the same upper bound for $\langle u, v_i\rangle - (N/M)\langle \Phi u, \Phi v_i\rangle$ and thus, with probability at least $1-\delta$,

$$\left|\frac{N}{M}\langle \Phi u, \Phi v_i\rangle - \langle u, v_i\rangle\right| \leq \frac{2\kappa_\delta}{2\sqrt{M}} \tag{31}$$

for $i = 1, 2, \ldots, |S|$. Finally, by substituting back in for u and v_i the theorem follows. □

We offer some brief comments. First, this bound is probably the best we can expect, since as the norm of either x or s_i grow, our bound ought to grow as well, and since in the special case of $s_i = 0$ our bound agrees with the fact that our estimate will be exact. Furthermore, note that the bound grows sub-linearly with the number of functions simultaneously estimated. While this is good, unfortunately the bound decays relatively slowly as a function of M, as illustrated in Figure 5. In a certain sense, this illustrates that estimation is a notably harder problem than detection or classification.

5 Related Work

Many of the functions we may wish to estimate are quite common. As such, the *data streaming* community (which is concerned with processing large streams of data using efficient algorithms) has analyzed many of them. In this setting one is often interested in estimating some function of the data stream, (such as an ℓ_p norm, a histogram, or a linear function) based on what are often called *sketches*, which can often be thought of as random projections. For a concise review of these results, see [27].

As a particularly relevant example, we note that for the case where $|S| = 1$, Theorem 4.1 is essentially the same (up to a constant) as a bound of [16] which uses Φ with 4-wise independent $\{+1, -1\}$-valued entries, yielding the result that with probability at least $1 - \delta$,

$$\left| \frac{1}{M} \langle \Phi x, \Phi s \rangle - \langle x, s \rangle \right| \leq \frac{2}{\sqrt{\delta}} \frac{\|x\|_2 \, \|s\|_2}{\sqrt{M}}. \tag{32}$$

The proof of this result in [16], while quite elementary, relies heavily on the special structure of the matrices considered. In light of Theorem 4.1, we can see that this bound actually holds for a significantly wider class of random matrices, and can be viewed a straightforward consequence of the same concentration of measure inequality that has proven useful for both compressed sensing and dimensionality reduction. Furthermore, our approach generalizes naturally to simultaneously estimating multiple linear functions of the data.

We expect that in the future such parallels between compressed estimation and sketches in data streaming will provide an avenue for the advancement of both areas.

6 Conclusion

The recently introduced theory of compressed sensing enables the reconstruction of sparse or compressible signals from a small set of nonadaptive, linear measurements. This has inspired the design of physical systems that directly implement similar measurement schemes. While most research has focused on the reconstruction of signals – typically relying on the assumption of sparsity – many (if not most) signal processing problems do not require a full reconstruction of the signal; we are often interested only in solving some sort of *detection* problem or in the *estimation* of some function of the data. We have shown that compressive measurements can be effective for a variety of detection, classification, and estimation problems. We have provided theoretical bounds along with experimental results. Our results place *no assumptions* on the signals being sparse or compressible. In the future we hope to provide a more detailed analysis of the classification setting and consider more general models, as well as consider detection, classification, and estimation settings that utilize more specific models – such as sparsity or manifold structure.

References

[1] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," 2005, Preprint.

[4] E. Candès and T. Tao, "Error correction via linear programming," *Found. of Comp. Math.*, 2005, Preprint.

[5] D. Donoho and Y. Tsaig, "Extensions of compressed sensing," 2004, Preprint.

[6] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 4036–4048, 2006.

[7] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, "Distributed compressed sensing," 2005, Preprint.

[8] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, K. F. Kelly, and R. G. Baraniuk, "A compressed sensing camera: New theory and an implementation using digital micromirrors," in *Proc. Computational Imaging IV at SPIE Electronic Imaging*, San Jose, CA, 2006, SPIE.

[9] M. B. Wakin, J. N. Laska, M. F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. F. Kelly, and R. G. Baraniuk, "An architecture for compressive imaging," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Atlanta, GA, 2006.

[10] J. A. Tropp, M. B. Wakin, M. F. Duarte, D. Baron, and R. G. Baraniuk, "Random filters for compressive sampling and reconstruction," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 2006.

[11] S. Kirolos, J. Laska, M. Wakin, M. Duarte, D. Baron, T. Ragheb, Y. Massoud, and R. Baraniuk, "Analog-to-information conversion via random demodulation," in *Proc. IEEE Dallas Circuits and Systems Workshop (DCAS)*, Dallas, TX, 2006.

[12] J. Laska, S. Kirolos, Y. Massoud, R. Baraniuk, A. Gilbert, M. Iwen, and M. Strauss, "Random sampling for analog-to-information conversion of wideband signals," in *Proc. IEEE Dallas Circuits and Systems Workshop (DCAS)*, Dallas, TX, 2006.

[13] J. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," 2005, Preprint.

[14] M. F. Duarte, M. A. Davenport, M. B. Wakin, and R. G. Baraniuk, "Sparse signal detection from incoherent projections," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, 2006, Toulouse, France.

[15] R. Castro, J. Haupt, and R. Nowak, "Active learning versus compressive sampling," in *Proc. Defense and Security Symp.*, Orlando, FL, 2006, SPIE.

[16] N. Alon, P. Gibbons, Y. Matias, and M. Szegedy, "Tracking join and self-join sizes in limited storage," in *Proc. Symp. Principles of Database Systems (PODS)*, Philadelphia, PA, 1999.

[17] E. Candès and T. Tao, "Near optimal signal recovery from random projections and universal encoding strategies," 2004, Preprint.

[18] A. Cohen, W. Dahmen, and R. DeVore, “Compressed sensing and best k-term approximation,” 2006, Preprint.

[19] R. G. Baraniuk, M. A. Davenport, R. A. DeVore, and M. B. Wakin, “The Johnson-Lindenstrauss lemma meets compressed sensing,” 2006, Preprint.

[20] W. B Johnson and J. Lindenstrauss, “Extensions of Lipschitz mappings into a Hilbert space,” in *Proc. Conf. in Modern Analysis and Probability*, 1984, pp. 189–206.

[21] S. Dasgupta and A. Gupta, “An elementary proof of the Johnson-Lindenstrauss lemma,” Tech. Rep. TR-99-006, Berkeley, CA, 1999.

[22] D. Achlioptas, “Database-friendly random projections,” in *Proc. Symp. Principles of Database Systems*, 2001.

[23] R. G. Baraniuk and M. B. Wakin, “Random projections of smooth manifolds,” 2006, Preprint.

[24] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*, Prentice Hall, 1998.

[25] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, Addison-Wesley, 1991.

[26] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions, Volume 1*, Wiley, 1994.

[27] S. Muthukrishnan, *Data Streams: Algorithms and Applications*, now, 2005.